Segmentación lingüística de tuplas para el modelado de la traducción

Centre de Recerca TALP
Universitat Politècnica de Catalunya (UPC)
Campus Nord UPC. 08034-Barcelona
{agispert,canton}@gps.tsc.upc.edu

Resumen: La traducción automática estocástica basada en n-gramas se fundamenta en un modelo de lenguaje de n-gramas estándar de unidades bilingües (tuplas) para modelar el proceso de la traducción, cuya estimación requiere de una segmentación para cada par de frases paralelas del corpus de entrenamiento. Esto implica la toma de ciertas decisiones firmes en cuanto a segmentación en unidades de traducción se refiere, especialmente cuando una palabra no es alineada a ninguna otra del otro idioma. En esta comunicación se presenta un estudio de esta situación, comparando técnicas de segmentación ya propuestas en dos tareas de traducción independientes: la tarea de gran vocabulario definida por el corpus de los debates de Parlamento Europeo entre inglés y español, y una tarea de tamaño mucho más reducido de expresiones turísticas entre el árabe y el inglés. Además, se propone una técnica de segmentación nueva que incorpora información lingüística, obteniendo mejores resultados en todas las tareas.
Palabras clave: traducción estocástica mediante n-gramas, segmentación en tuplas, modelo de traducción

Abstract: Ngram-based Statistical Machine Translation relies on a standard Ngram language model of tuples to estimate the translation process. In training, this translation model requires a segmentation of each parallel sentence, which involves taking a hard decision on tuple segmentation when a word is not linked during word alignment. This is especially critical when this word appears in the target language, as this hard decision is compulsory. In this paper we present a thorough study of this situation, comparing for the first time each of the proposed techniques in two independent tasks, namely English–Spanish European Parliament Proceedings large-vocabulary task and Arabic–English Basic Travel Expressions small-data task. In the face of this comparison, we present a novel segmentation technique which incorporates linguistic information. Results obtained in both tasks outperform all previous techniques.
Keywords: Ngram-based statistical machine translation, tuple segmentation, translation model

1. Introducción

Los sistemas de traducción estocástica basados en n-gramas han demostrado ser una alternativa viable al enfoque basado en *phrases*, obteniendo sistemáticamente resultados del estado del arte en sucesivas evaluaciones (Koehn y Monz, 2005; Eck y Hori, 2005). Su principal diferencia radica en la estimación del modelo de traducción por medio de un modelo de lenguaje de n-gramas, definido en el bilenguaje expresado por las tuplas (Mariño et al., 2005).

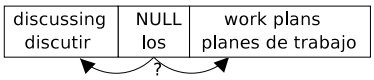
Según se muestra en la literatura, las tuplas son unidades que contienen *una o más* palabras del idioma fuente y *una o más* palabras del idioma destino, inluyendo el token NULO o palabra vacía, que en realidad no es ninguna palabra. Este modelo tiene sus orígenes en la traducción estocástica utilizando transductores de estados finitos (Vidal, 1997; Casacuberta, Vidal, y Vilar, 2002), cuya formulación matemática expresa que se busca aquella frase destino 'd' que maximiza:

$$\prod_{n=1}^{N} p((d,f)_n|(d,f)_{n-x+1},...,(d,f)_{n-1}) \quad (1)$$

donde $(d,f)_n$ se refiere a la n-*ésima* tupla de un determinado par de frases destino y fuente. Para estimar los parámetros de este modelo dado un corpus paralelo y su alineamiento a nivel de palabras, la fase de entrenamiento, a diferencia de la utilizada en los enfoques basados en *phrases*, requiere de una segmentación **única** de cada par de frases paralelas en una secuencia de tuplas, de forma que se respete el orden natural de ambos

idiomas, como se describe en (Crego, Mariño, y de Gispert, 2004).

Sin embargo, el algoritmo de extracción o generación de tuplas introducido en (Crego, Mariño, y de Gispert, 2004)) define un conjunto único de tuplas excepto cuando la tupla extraída a partir del alineamiento no contiene ninguna palabra fuente (o contiene el token NULO). Para poder reutilizar dicha tupla en el momento de traducir nuevas frases, sería necesario permitir al algoritmo de decodificación de la traducción que aceptara generar palabras destino sin cubrir ninguna palabra de la frase fuente a traducir. Ningún decodificador de traducción estocástica permite tal comportamiento, y por lo tanto, para estos casos se debe tomar una decisión firme referente a segmentación de unidades. Véase el siguiente ejemplo, en donde se ilustra la necesidad de decidir si se junta la tupla con fuente NULO a la tupla previa o a la siguiente.



La literatura presenta ejemplos de criterios para tomar dicha decisión de segmentación, que principalmente van desde simplemente juntar todas las tuplas con fuente NULO a la tupla siguiente (o anterior) de forma determinista, hasta comparar las probabilidades del modelo IBM 1 asociadas a las tuplas resultantes de las dos segmentaciones resultantes (Crego, Mariño, y Gispert, 2005).

Sin embargo, el impacto que esta decisión de segmentación pueda tener sobre la calidad de la traducción (bondad del modelo de traducción) no ha sido estudiada, ni tampoco se han comparado los distintos métodos propuestos. En esta contribución se pretende realizar dicha comparación, explorando qué grado de importancia tiene esta decisión a la hora de estimar un modelo de traducción basado en n-gramas y de traducir un conjunto test. Además, se propone un nuevo criterio de segmentación que utiliza información lingüística a través de la entropía de las etiquetas morfológicas (o Part-Of-Speech) para, indirectamente, redcir la entropía del modelo. La comparativa se realiza en dos tareas diferenciadas en tamaño y par de lenguas implicadas, como son una tarea inglés–español (y viceversa) de gran vocabulario, y una tarea árabe–inglés de tamaño más reducido. Por último, también se estudia el efecto de tomar las mismas decisiones para los tokens NULO que aparecen en el idioma destino, y su impacto en traducción.

El artículo está organizado de la siguiente manera. La sección 2 repasa las estrategias de segmentación existentes y propone un nuevo criterio lingüístico, analizando las ventajas y desventajas de cada enfoque. La sección 3 presenta el trabajo experimental realizado, y por último, la sección 4 presenta las conclusiones del estudio, y la sección 5 proporciona ideas para el trabajo futuro, con el objectivo de mejorar el modelo de traducción basado en n-gramas.

2. Criterios de Segmentación de Tuplas

En referencia al modelo de traducción basado en n-gramas, parece evidente que la estrategia ideal para la segmentación de tuplas debería tomar una decisión global basada en la segmentación de todas las unidades con fuente NULO, intentando obtener aquel conjunto de tuplas y n-gramas que representase mejor el universo no observado. Sin embargo, no existe ningún algoritmo viable que sea capaz de realizar dichos cálculos en un tiempo razonable, puesto que esto implica el volver a estimar el modelo para cada alternativa de segmentación.

Hasta ahora, sólo se han propuesto dos estrategias de segmentación para resolver el problema de las unidades con fuente NULO, que se presentan a continuación, conjuntamente con las nueva propuesta siguiendo un enfoque más lingüístico.

2.1. Determinista a la siguiente

Muy pragmático y simple, este enfoque consiste en juntar todas las palabras destino pertenecientes a una tupla con fuente NULO, *a la siguiente* unidad de traducción (salvo cuando es la última de la frase, en cuyo caso va a la anterior), como se introdujo en (de Gispert y Mariño, 2004). Aparte de la simplicidad y eficiencia extrema, no encontramos otra ventaja de este enfoque, que no sigue ningún criterio lingüístico ni estocástico.

2.2. Peso del modelo IBM 1

Las probabilidades del modelo de IBM 1 proporcionan un lexicón probabilístico entre pares de palabras fuente y destino, independiente de su posición en la frase (véase (Brown et al., 1993) para detalles sobre estos modelos). Esta información puede utilizarse para proporcionar un peso y comparar las tuplas resultantes de las dos segmentaciones posibles, como se introdujo en (Crego, Mariño, y Gispert, 2005). Dicho peso se define para cada tupla como:

$$\frac{1}{I}\prod_{j=1}^{J}\sum_{i=0}^{I} p_{IBM1}(d^i|f^j)p_{IBM1'}(f^j|d^i) \quad (2)$$

donde f y d representan las partes fuente y destino de la tupla, I y J sus respectivos tamaños en número de palabras, e $IBM1'$ representa las probabilidades del modelo IBM 1 estimado en la dirección opuesta (de destino a fuente).

Aunque este enfoque es atractivo en el hecho de que utiliza información bilingüe en la toma de decisiones de segmentación, la observación de las tuplas resultantes en estas situaciones revela una conclusión distinta. Muchas de las palabras destino pertenecientes a tuplas con fuente NULO son artículos, preposiciones, conjunciones y otras partículas cuya principal función es asegurar la cohesión gramatical de la frase destino, complementando otras palabras más informativas semánticamente. Esto hace que sus probabilidades de traducción a las palabras del otro idioma tengan poco sentido, ya que en muchos casos no tienen una palabra correspondiente en la traducción.

2.3. Entropía de la distribución de POS

Por otro lado, y desde un punto de vista más lingüístico, se puede ver el problema de la segmentación de tuplas alrededor de las palabras asociadas al token NULO como una decisión monolingüe referente a si una determinada palabra del idioma destino está más vinculada lingüisticamente con la palabra siguiente o con la anterior.

De forma intuitiva, podemos esperar que un buen criterio de segmentación será aquél que preserve las unidades conexas gramaticalmente (como por ejemplo los artículos que preceden a un determinado sustantivo) en la misma tupla, puesto que esto lleva a una simplificación de la tarea de traducción. Por contra, separar unidades lingüísticas en distintas tuplas probablemente provocaría un incremento del vocabulario de tuplas, una mayor escasez de datos y un modelo de n-gramas de traducción más pobre.

En esta línea de razonamiento, se propone tomar la decisión de segmentación de acuerdo con las entropías *anterior y posterior* de las distribuciones de etiquetas morfológicas (o Part-Of-Speech), que se definen en función del contexto de palabras. En concreto, dada la siguiente secuencia de 3 tuplas:

$$\begin{array}{ccc} < ...f_j > & \text{NULO} & < f_{j+1}... > \\ | & | & | \\ < ...d_{i-1} > & d_i & < d_{i+1}... > \end{array}$$

donde f_j es la palabra en la posición j de la frase fuente, y equivalentemente d_i es la palabra en la posición i de la frase destino, podemos definir una entropía 'posterior' de la distribución de POS en la posición $i+1$ dadas las palabras previas (d_{i-1}, d_i) como se expresa en la ecuación 3:

$$H^p_{POS} = -\sum_{POS} p^p_{POS} \log p^p_{POS} \quad (3)$$

donde

$$p^p_{POS} = \frac{N(d_{i-1}, d_i, POS_{i+1})}{\sum_{POS'} N(d_{i-1}, d_i, POS'_{i+1})} \quad (4)$$

es la probabilidad de observar una determinada etiqueta morfológica *siguiendo* a la secuencia de palabras definida por d_i and d_{i+1}, estimada por frecuencia relativa.

Equivalentemente, podemos definir una entropía 'anterior' de la distribución de POS en la posición $i-1$ dadas las palabras siguiente (d_i, d_{i+1}) como es la ecuación 5:

$$H^a_{POS} = -\sum_{POS} p^a_{POS} \log p^a_{POS} \quad (5)$$

donde

$$p^a_{POS} = \frac{N(POS_{i-1}, d_i, d_{i+1})}{\sum_{POS'} N(POS'_{i-1}, d_i, d_{i+1})} \quad (6)$$

es la probabilidad de observar una determinada etiqueta morfológica *precediendo* a la secuencia de palabras definida por d_{i-1} y d_i.

hay	NULO	ninguna
|	|	|
there	are	no
$H^p_{POS}(there, are, ***) = 0{,}83$		
$H^a_{POS}(***, are, no) = 0{,}62$		

Cuadro 1: *Example of H^f_{POS} and H^b_{POS} entropies.*

Una vez calculadas dichas entropías, podemos tomar la decisión de segmentación eligiendo el caso con mayor entropía de POS. Esto se justifica debido a que, si $H^p_{POS} > H^a_{POS}$, hemos

Source	Target	siguiente	pesoIBM1	entropPOS
We	Nos	We — Nos		
are pleased	alegramos	are pleased — alegramos		
at	NULO	at—NULO	at—de	at—de
NULO	**de**	this—de esta		
this	esta		this—esta	this—esta
visit	visita	visit — visita		
which	que	which—que	which—que se	which—que
NULO	**se**	reflects—se enmarca		reflects—se enmarca en
reflects	enmarca		reflects—enmarca	
NULO	**en**	the—en la	the—en la	
the	la			the—la
cooperation	cooperación	cooperation — cooperación		
between	entre	between — entre		
parliaments	parlamentos	parliaments — parlamentos		
in	NULO	in—NULO	in—NULO	in—de
NULO	**de**	the—de la	the—de la	
the	la			the—la
Union	Unión	Union — Unión		

Cuadro 2: *Ejemplo de las decisiones de segmentación tomadas alrededor de las palabras alineadas a NULO por los distintos criterios en una frase del corpus inglés–español.*

observado la primera secuencia de palabras compuesta por (d_{i-1}, d_i) en más situaciones gramaticalmente diferentes que la otra secuencia compuesta por (d_i, d_{i+1}). Por tanto, podemos concluir que d_{i-1} y d_i están más vinculadas que d_i y d_{i+1}, y por lo tanto deberían pertenecer a la misma tupla de traducción. De forma análoga, se puede concluir lo contrario si $H^p_{POS} < H^a_{POS}$.

Para ilustrar esta idea, véase el ejemplo del cuadro 1, donde la entropía posterior de las palabras inglesas 'there are' es superior a la entropía anterior del par 'are no', lo que indica que 'there are' debe formar parte de la misma tupla.

El cuadro 2 muestra una frase de ejemplo de inglés a español, donde los enlaces originales (procedentes del alineamiento de palabras) se muestran en las primeras dos columnas, mientras que se comparan las tres estrategias de segmentación mencionadas en las siguientes columnas. Como se puede observar, las tuplas resultantes en el caso basado en entropía de las etiquetas POS son muy cercanas a lo esperable desde un punto de vista intuitivo.

2.4. El NULO en el destino

Mientras que la decisión de segmentación es obligatoria cuando una palabra destino no está alineada (lo que equivale a estar alineada al token NULO), no sucede lo mismo cuando la unidad no alineada está en el idioma fuente. En este caso, se generan tuplas que traducen a NULO y se aceptan como parte del vocabulario del modelo de traducción (a diferencia del enfoque basado en phrases (Zens, Och, y Ney, 2002), donde no existen los tokens NULO y por lo tanto se posponen las decisiones de segmentación al tiempo de decodificación de la traducción).

Sin embargo, se puede pensar en aplicar el mismo criterio de segmentación para las palabras fuente alineadas a NULO, con el objetivo de estudiar cuál es la aportación de dichas unidades a la calidad de la traducción, y si eliminándolas del modelo puede el sistema incurrir en menos errores de omisiones. En el trabajo experimental también se ha abordado esta situación, como se verá en la sección siguiente.

3. Trabajo experimental

A fin de comparar las distintas estrategias de segmentación y de evaluar su impacto en la calidad de la traducción, se han realizado experimentos utilizando dos corpora paralelos, que difieren en tamaño y par de lenguas implicadas. Por un lado, se ha utilizado un corpus inglés–español de gran vocabulario, correspondiente a las transcripciones de los debates del Parlamento Europeo desde 1996 hasta el 2004. Por otro lado, un corpus árabe–inglés de vocabulario reducido, que contiene una parte del denominado Basic Travel Expressions Corpus (BTEC). Los textos ingleses han sido etiquetados morfológicamente utilizando la herramienta *TnT* tagger[1], mientras que el

[1]Disponible en www.coli.uni-saarland.de/∼thorsten/tnt

texto español ha sido etiquetado mediante el paquete de análisis morfosintáctico *FreeLing*[2].

En el cuadro 3 se muestran las estadísticas de ambos corpora, donde cabe destacar la notable diferencia en tamaño.

3.1. Estadísticas de tuplas

Para el conjunto de entrenamiento (entr.), el cuadro 3 también muestra el número de tuplas extraídas a partir del alineamiento de palabras[3], así como el porcentaje de tuplas con NULO en alguna de sus partes. Como es esperable, este porcentaje es superior para el inglés (14.5 %), dado que el español contiene más palabras que, por lo tanto, no tendrán una correspondencia directa en inglés.

Dado que la traducción estocástica basada en n-gramas precisa de un modelo sin NULOs en el idioma fuente, en la dirección ing→esp se debe tomar una decisión firme para el 14.5 % de las tuplas, mientras que en la dirección opuesta, sólo el 11.7 % de las tuplas debe resegmentarse. Por tanto, cabe esperar un mayor impacto de las estrategias de segmentación para el primer caso.

3.2. Resultados del modelo de traducción

		BLEU	mWER	NIST
I→E	siguiente	0.4215	43.98	9.22
	pesoIBM1	0.4221	43.60	9.19
	entropPOS	**0.4325**	**43.48**	**9.30**
	destNULO	0.4249	44.47	9.21
	destNULOpos	0.4313	43.75	9.29
E→I	siguiente	0.4661	39.37	9.86
	pesoIBM1	0.4698	38.73	9.91
	entropPOS	**0.4756**	**38.64**	**9.95**
	destNULO	0.4728	39.23	9.91
	destNULOpos	0.4733	38.78	9.93
A→I	siguiente	0.3684	41.80	7.16
	pesoIBM1	0.3656	41.94	7.14
	entropPOS	**0.3691**	41.91	**7.17**

Cuadro 4: *Resultados del modelo de traducción para cada estrategia de segmentación. 'I' significa inglés, 'E' español y 'A' árabe.*

En el cuadro 4 se muestra una comparación de resultados del modelo de traducción de n-gramas para cada tarea, en las filas 'siguiente', 'pesoIBM1' y 'entropPOS', en referencia a las estrategias de segmentación presentadas en la sección anterior.

En cuanto a las tareas de gran vocabulario, la segmentación lingüística propuesta obtiene prestaciones significativamente mejores que las demás estrategias, especialmente en la dirección I→E. Este resultado es coherente con el hecho de que el español genera más palabras que el inglés y, por lo tanto, el porcentaje de tuplas con NULO en el fuente es superior (como se mencionó en la sección 3.1).

En la dirección E→I, a pesar de que el impacto de cambiar el criterio de segmentación es menor, la mejora producida por el enfoque entropPOS es significativa. En la tarea de vocabulario reducido A→I, las diferencias son menos significativas, en correlación con el hecho de que sólo el 7 % de las tuplas contienen NULO en la parte árabe, comparado con el 14 % de la tarea I→E (véase cuadro 3).

Cabe remarcar que, mientras que la estrategia pesoIBM1 proporciona mejores prestaciones que el criterio 'siguiente' para las tareas de gran vocabulario, el resultado es opuesto en la tarea A→I. En cambio, el enfoque entropPOS se muestra más robusto a un cambio de tarea, obteniendo resultados óptimos en todos los casos.

3.3. Eliminación de los NULO en destino

Aplicando las ideas introducidas en la sección 2.4, el cuadro 4 también presenta los resultados al aplicar el mejor criterio de segmentación (entropPOS) para eliminar las tuplas con NULO en la parte *destino*, como se muestra en las filas 'destNULO' y 'destNULOpos' para las tareas inglés–español. El primer caso se refiere a aplicar el criterio para eliminar todas estas tuplas, mientras que el segundo aplica únicamente a aquellas tuplas que contienen un sustantivo, adjetivo o verbo en su parte fuente. El objetivo es evitar los errores por omisión de palabras con mayor contenido semántico en la traducción.

Sin embargo, los resultados muestran que ninguna de estas estrategias proporciona una calidad mejor. A diferencia de los NULOs en el fuente, los NULOs en el destino parecen ser para el modelo de n-gramas un mecanismo útil para aprender contextos de traducción y aportan mejoras en todas las direcciones probadas. La misma conclusión aplica para el caso 'destNULOpos', aunque se observa una ligera mejora de la calidad.

[2]Disponible en http://garraf.epsevg.upc.es/freeling

[3]Alineamiento unión de los alineados en las direcciones f→d y d→f, obtenidos con la herramienta GIZA++, disponible en www.fjoch.com

		Parlamento Europeo		Basic Travel Expressions	
		español	inglés	árabe	inglés
entr.	Frases	1223398		20000	
	Palabras	34963601	33374308	180477	189160
	Vocabulario	151476	104826	15956	7169
	long. media	28.6	27.3	9.0	9.5
	Tuplas	20032806		122176	
	Tuplas con NULO	11.7 %	14.5 %	7.0 %	7.2 %
des.	Frases	504		506	—
	Palabras	15415	15331	3632	—
	Vocabulario	2735	2300	1179	—
	Palabras desc.	22	20	196	—
	Referencias	3		—	16
test	Frases	840	1094	1006	—
	Palabras	22753	26876	7217	—
	Vocabulario	4085	3975	1884	—
	Palabras desc.	44	113	356	—
	Referencias	2		—	16

Cuadro 3: *Estadísticas de los dos corpora paralelos utilizados, incluyendo número de frases y palabras, talla del vocabulario, longitud media de las frases y, para los conjuntos de desarrollo y test, número de palabras desconocidas y de traducciones referencia utilizadas para evaluar.*

		vcb tup	% 1–2–3gramas	lon tup	NULOs
E→S	siguiente	2110085	17.6 – 44.4 – 38.0	1.157-1.096	3119
	pesoIBM1	2035523	18.0 – 44.7 – 37.3	1.157-1.090	2466
	POSentropy	2084640	17.8 – 44.3 – 37.9	1.156-1.106	2282
	destNULO	2347743	23.2 – 45.1 – 31.7	1.253-1.190	0
	destNULOpos	2178470	19.0 – 44.5 – 36.5	1.180-1.139	1625
S→E	siguiente	2149595	14.1 – 41.5 – 44.4	1.135-1.064	2761
	pesoIBM1	2080171	14.2 – 41.4 – 44.4	1.131-1.054	2318
	POSentropy	2109351	14.2 – 41.5 – 44.3	1.134-1.064	2194
	destNULO	2421446	19.9 – 44.1 – 36.0	1.260-1.224	0
	destNULOpos	2164076	14.7 – 41.6 – 43.7	1.143-1.075	1977

Cuadro 5: *Vocabulario de tuplas y estadísticas de n-gramas de traducción para cada segmentación.*

3.4. N-gramas de traducción

Para comprender mejor estos resultados, el cuadro 5 muestra la talla del vocabulario de tuplas obtenido para cada segmentación (vcb tup), así como estadísticas relevantes de la salida traducida, como el porcentaje de tuplas del test que han sido observadas como 1-gramas, 2-gramas y 3-gramas en el entrenamiento, la longitud media de la tupla (para las partes fuente y destino por separado), y el número de tuplas con NULO en el destino (para la traducción generada).

En cuanto a vocabulario, el criterio 'siguiente' produce la mayor talla en entrenamiento, seguido del entropPOS y por último del pesoIBM1. Al eliminar los NULOs en destino, la talla del vocabulario incrementa notablemente.

En I→E, observamos que la traducción con los criterios 'siguiente' y 'entropPOS' tiende a utilizar más 3-gramas que con 'pesoIBM1', lo que puede explicarse por su consistencia a la hora de tomar decisiones (siempre segmentan igual para las mismas palabras destino implicadas), mientras que pesoIBM1 depende de información bilingüe y es más variable.

Sin embargo, el hecho de utilizar más 3-gramas en traducción no está correlacionado directamente con las medidas de calidad, y hay que tener en cuenta el número de tuplas con NULO en destino, que es marcadamente superior para el caso 'siguiente'. Esto indica que en este caso se están encadenando muchos 3-gramas de NULOs en destino, lo que no aporta la mejor traducción.

En el caso 'pesoIBM1' y especialmente 'entropPOS', el número de tuplas con NULO en el destino es mucho inferior.

Aunque esto parece ser positivo para la traducción, cuando eliminamos completa o parcialmente los NULOs del destino (destNULO y destNULOpos), la longitud media de tupla aumenta, no sólo en el fuente sino también en el idioma destino, y el modelo pierde contexto de tuplas, cayendo mucho más al 1-grama. Esto tiene un efecto negativo en la calidad de la traducción.

Por lo tanto, podemos concluir que la mejor relación entre mayor contexto de tuplas (n-grama largo) y menor cantidad de tuplas a NULO se da con la segmentación entropPOS propuesta.

Las diferencias son mucho menores en la dirección I→E, aunque se observa la misma tendencia en el número de tuplas con NULO en destino, y las conclusiones son análogas.

3.5. Impacto absoluto

A fin de estudiar el impacto absoluto de la toma de decisiones de segmentación, hemos definido como el peor caso la toma aleatoria y se ha evaluado los resultados de traducción, como se muestra en el cuadro 6, donde 'aleat' es el resultado mediana de 5 experimentos.

		BLEU	mWER	NIST
I→E	aleat	0.4202	43.80	9.17
E→I	aleat	0.4707	38.60	9.92
A→I	aleat	0.2758	50.74	5.78

Cuadro 6: *Resultados para el peor caso.*

Sorprendentemente, las estrategias 'siguiente' y 'pesoIBM1' obtienen resultados similares al peor caso, e incluso peores para E→I. Teniendo en cuenta la baja significación estadística de realizar sólo 5 experimentos, la conclusión cualitativa es que ninguna de estas estrategias mejoran significativamente el caso aleatorio.

En cambio, en la tarea A→I, probablemente debido al tamaño del corpus, la estrategia aleatoria provoca una mayor escasez de datos y un resultado de traducción muy pobre.

3.6. Resultados modelo de traducción + características

Para mayor evaluación del impacto de la segmentación, se ha combinado de forma log-linear el modelo de traducción de n-gramas con 4 modelos adicionales: dos modelos léxicos basados en probabilidades del modelo IBM 1, un modelo de lenguaje destino y una bonificación constante a la generación de palabras, cuyos pesos se han optimizado según el BLEU obtenido en el conjunto de desarrollo (de forma similar a como se realiza en (Mariño et al., 2005)). El cuadro 7 muestra los resultados de traducción para las dos mejores segmentaciones en cada tarea. Como se puede observar, la mejora proporcionada por la estrategia entropPOS es prácticamente compensada por los modelos adicionales, especialmente en E→I.

		BLEU	mWER	NIST
I→E	pesoIBM1	0.4714	40.22	9.83
	entropPOS	**0.4744**	40.56	**9.85**
E→I	pesoIBM1	0.5470	34.41	10.74
	entropPOS	0.5466	34.44	10.72
A→I	alwaysNEXT	0.3974	40.16	7.23
	entropPOS	**0.4024**	**40.05**	**7.39**

Cuadro 7: *Resultados del modelo de traducción con modelos adicionales para cada segmentación.*

En las tareas inglés–español de gran vocabulario, los modelos del destino y léxicos proporcionan robustez al sistema penalizando las tuplas con mala segmentación, o por lo menos su concatenación en n-gramas de traducción. Aun así, la segmentación propuesta obtiene resultados ligeramente superiores para la tarea I→E.

Sin embargo, las tareas de poco vocabulario son mucho más sensibles a la segmentación incluso cuando se combina el modelo de traducción con modelos adicionales, y las mejoras son más significativas (ver caso A→I).

4. Conclusiones

Esta contribución estudia con detalle la segmentación y extracción de tuplas, un proceso clave en el entrenamiento de sistemas de traducción estocástica basados en n-gramas. Además de revisar, estudiar y comparar los criterios de segmentación previamente presentados, se propone una nueva estrategia basada en la distribución de etiquetas morfológicas (Part-Of-Speech).

Las principales conclusiones de este trabajo son:

- Las prestaciones del modelo de traducción están afectadas significativamente por la segmentación de tuplas, cuyo impacto depende del par de lenguas implicadas y del tamaño del corpus utilizado, siendo mayor cuando aumenta el porcentaje de tuplas con NULO en el fuente

- Las estrategias de segmentación ya propuestas no superan notablemente el caso aleatorio para tareas de gran vocabulario, mientras que la estrategia lingüística propuesta es significativamente mejor
- Para tareas de vocabulario reducido, la segmentación aleatoria empobrece mucho el modelo, mientras que la estrategia entropPOS obtiene resultados óptimos, comportándose de forma robusta al cambio de tarea de traducción
- En cuanto a los NULOs en la parte destino, proporcionan contexto útil al modelo de traducción y eliminarlos a través de la resegmentación no es beneficioso para la calidad de la traduccción
- Cuando el modelo de traducción se combina con otras funciones características, el impacto directo de la segmentación es menor (las demás funciones pueden compensar parcialmente una mala segmentación de las unidades) para las tareas de gran vocabulario

5. Trabajo futuro

Dado que la estrategia propuesta requiere de un etiquetado morfológico, una posible solución ante la falta de dicha herramienta puede ser la clasificación automática de palabras, como se propone en Rapp (2005). En dirección opuesta, si se dispone de herramientas de chunking (o etiquetado sintáctico superficial), sería interesante investigar formas de utilizar dicha información a la hora de segmentar unidades de traducción.

Otra línea de investigación futura se refiere a los NULOs en destino. A pesar de que eliminar total o parcialmente dichas unidades del modelo de traducción no mejora resultados, parece que la mejor estrategia es aquella que utiliza menos unidades a NULO en la traducción. En un futuro se pretende realizar un estudio profundo de esta aparente paradoja, a fin de esclarecer cómo utiliza el modelo estas unidades y si es posible mejorar la traducción eliminando algunas.

Agradecimientos

Este trabajo ha sido cofinanciado por el proyecto TC-STAR (Unión Europea, FP6-506738), la Generalitat de Catalunya y el Fondo Social Europeo.

Bibliografía

Brown, P., S. Della Pietra, V. Della Pietra, y R. Mercer. 1993. The mathematics of statistical machine translation. *Computational Linguistics*, 19(2):263–311.

Casacuberta, F., E. Vidal, y J.M. Vilar. 2002. Architectures for speech-to-speech translation using finite-state models. *Proceedings of the Workshop on Speech-to-Speech Translation: Algorithms and Systems*, páginas 39–44, July.

Crego, J. M., J. Mariño, y A. de Gispert. 2004. Finite-state-based and phrase-based statistical machine translation. *Proc. of the 8th Int. Conf. on Spoken Language Processing, ICSLP'04*, páginas 37–40, October.

Crego, J. M., J. Mariño, y A. Gispert. 2005. TALP: The UPC tuple-based SMT system. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, páginas 191–198, October.

de Gispert, A. y J. Mariño. 2004. TALP: Xgram-based Spoken Language Translation System. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, páginas 85–90, October.

Eck, M. y Ch. Hori. 2005. Overview of the IWSLT 2005 Evaluation Campaign. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'05*, páginas 11–32, October.

Koehn, P. y C. Monz. 2005. Shared task: Statistical Machine Translation between European Languages. *Proc. of the ACL Workshop on Building and Using Parallel Texts (ACL'05)*, páginas 119–124, June.

Mariño, J.B., R Banchs, J.M. Crego, A. de Gispert, P. Lambert, M. R. Costa-jussà, y J.A.R. Fonollosa. 2005. Bilingual N–gram statistical machine translation. *Proc. of the MT Summit X*, páginas 275–282, September.

Rapp, R. 2005. A practical solution to the problem of automatic part-of-speech induction from text. En *Proc. of 43rd Annual Meeting of the ACL (Companion Volume)*, páginas 77–80, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Vidal, E. 1997. Finite-state speech-to-speech translation. *Proc. of 1997 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, páginas 111–114, April.

Zens, R., F.J. Och, y H. Ney. 2002. Phrase-based statistical machine translation. En M. Jarke J. Koehler, y G. Lakemeyer, editores, *KI - 2002: Advances in artificial intelligence*, volumen LNAI 2479. Springer Verlag, September, páginas 18–32.